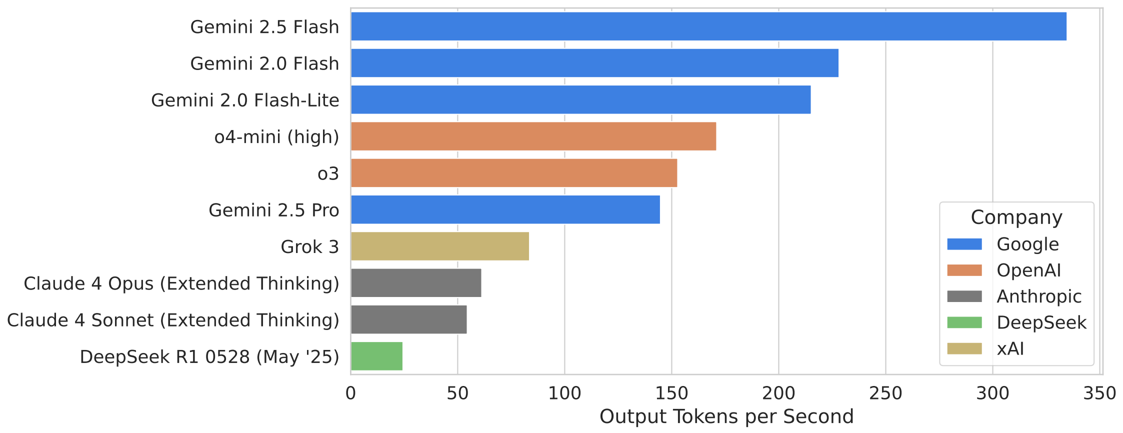

Figure 2 | Number of output tokens generated per second (after the first chunk has been received from the API) for different models. Source: `ArtificialAnalysis.ai`, imported on 2025-06-15.

we also utilized new methods for improved data quality for both filtering, and deduplication. Our post-training dataset, like Gemini 1.5, consists of instruction tuning data that is carefully collected and vetted. It is a collection of multimodal data with paired instructions and responses, in addition to human preference and tool-use data.

2.3. Training Infrastructure

This model family is the first to be trained on TPUv5p architecture. We employed synchronous data-parallel training to parallelise over multiple 8960-chip pods of Google's TPUv5p accelerators, distributed across multiple datacenters.

The main advances in software pre-training infrastructure compared with Gemini 1.5 were related to elasticity and mitigation of SDC (Silent Data Corruption) errors:

1. **Slice-Granularity Elasticity**: Our system now automatically continues training with fewer "slices" of TPU chips when there is a localized failure, and this reconfiguration results in tens of seconds of lost training time per interruption, compared with the 10 or more minute delay waiting for healthy machines to be rescheduled without elasticity; the system continues training at around 97% throughput while the failed slice is recovering. At the scale of this training run we see interruptions from hardware failures multiple times per hour, but our fault tolerance machinery is designed to tolerate the higher failure rates expected at much larger scales.
2. **Split-Phase SDC Detection**: On previous large-scale runs it could take many hours to detect and localize machines with SDC errors, requiring both downtime while debugging, and roll-back/replay of a large number of potentially corrupt training steps. We now use lightweight deterministic replay to immediately repeat any step with suspicious metrics, and compare per-device intermediate checksums to localize the root cause of any data corruption. Empirically, accelerators that start to exhibit intermittent SDCs are identified within a few minutes, and quickly excluded from the job. During this run, around 0.25% of steps were replayed due to suspected SDCs and 6% of these replays turned out to be genuine hardware corruption.

Both of the above techniques were relatively simple to implement due to the single-controller design of the Pathways system (Barham et al., 2022), which allows all accelerators to be coordinated from a single python program with a global view of the system state. The controller can make use of

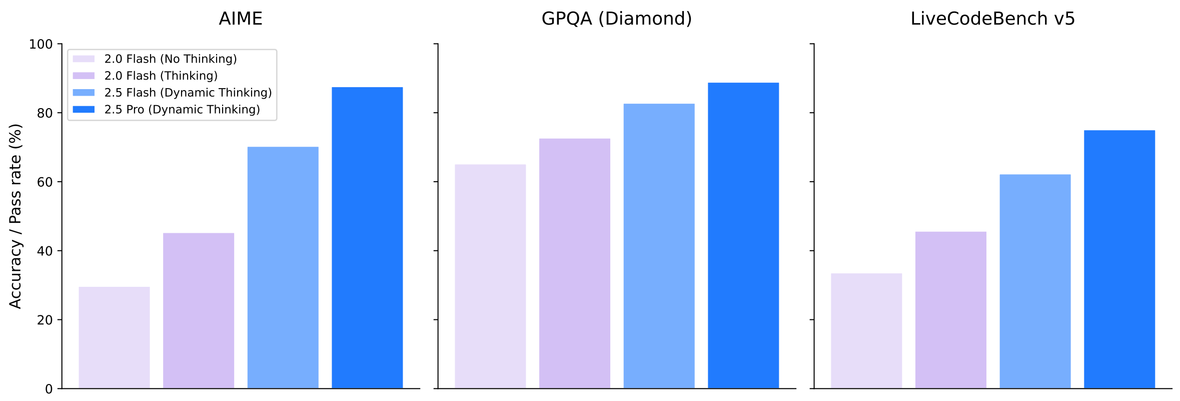

Figure 3 | Impact of "Thinking" on Gemini's performance on AIME 2025 (Balunović et al., 2025), LiveCodeBench (corresponding to 10/05/2024 - 01/04/2025 in the UI) (Jain et al., 2024) and GPQA diamond (Rein et al., 2024) benchmarks.

parallel 'remote python' operations on TPU workers to monitor training metrics, track performance stragglers, and root-cause SDC errors.

Overall during the run, 93.4% of the time was spent performing TPU computations; the remainder was approximately spent half in elastic reconfigurations, and half in rare tail cases where elasticity failed. Around 4.5% of the computed steps were replays or rollbacks for model debugging interventions.

2.4. Post-training

Since the initial announcement of Gemini 1.5, significant advancements have been made in our post-training methodologies, driven by a consistent focus on data quality across the Supervised Fine-Tuning (SFT), Reward Modeling (RM), and Reinforcement Learning (RL) stages. A key focus has been leveraging the model itself to assist in these processes, enabling more efficient and nuanced quality control.

Furthermore, we have increased the training compute allocated to RL, allowing deeper exploration and refinement of model behaviors. This has been coupled with a focus on verifiable rewards and model-based generative rewards to provide more sophisticated and scalable feedback signals. Algorithmic changes to the RL process have also improved stability during longer training. These advancements have enabled Gemini 2.5 to learn from more diverse and complex RL environments, including those requiring multi-step actions and tool use. The combination of these improvements in data quality, increased compute, algorithmic enhancements, and expanded capabilities has contributed to across-the-board performance gains (as described in Section 3) , notably reflected in the significant increase in the model's LMArena Elo scores, with both Gemini 2.5 Flash and Pro gaining more than 110 points over their Gemini 1.5 counterparts (122 for Gemini 2.5 Pro and 111 for Gemini 2.5 Flash, see Figure 1), along with significant improvements on several other frontier benchmarks.

2.5. Thinking

Past Gemini models produce an answer immediately following a user query. This constrains the amount of inference-time compute (Thinking) that our models can spend reasoning over a problem. Gemini Thinking models are trained with Reinforcement Learning to use additional compute at inference time to arrive at more accurate answers. The resulting models are able to spend tens of

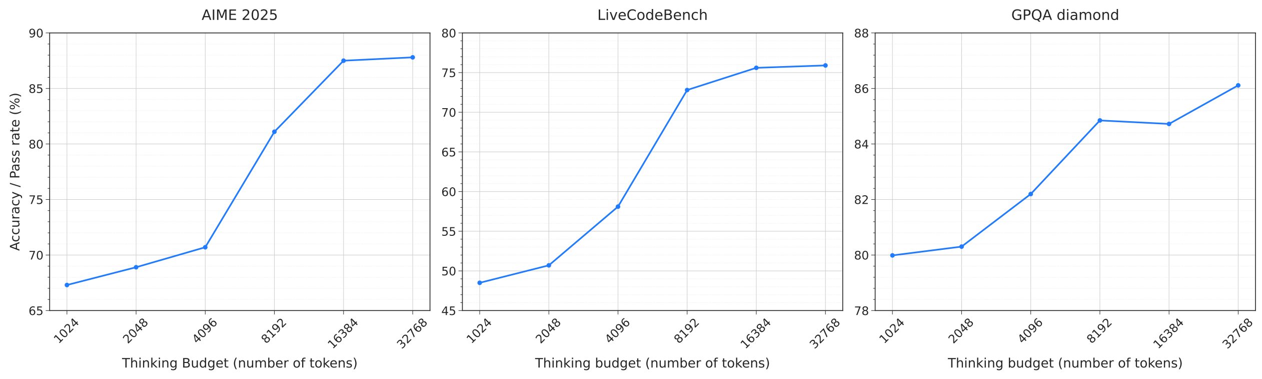

Figure 4 | Impact of thinking budget on performance on AIME 2025 (Balunović et al., 2025), LiveCodeBench (corresponding to 10/05/2024 - 01/04/2025 in the UI) (Jain et al., 2024) and GPQA diamond (Rein et al., 2024) benchmarks.

thousands of forward passes during a "thinking" stage, before responding to a question or query.

Our training recipe has evolved from the original experimental thinking model, Gemini 2.0 Flash Thinking (launched in December 2024), to the Gemini 2.5 Thinking series, which incorporates Thinking natively across all domains. The result is a single model that can achieve stronger reasoning performance across the board, and is able to scale up its performance further as a function of inference time (see Figure 3 for an example of the impact of Thinking).

We integrated Thinking with other Gemini capabilities, including native multimodal inputs (images, text, video, audio) and long context (1M+ tokens). For any of these capabilities, the model decides for itself how long to think before providing an answer. We also provide the ability to set a Thinking budget, constraining the model to respond within a desired number of tokens. This allows users to trade off performance with cost. To demonstrate this capability, we conducted experiments where we systematically varied the thinking budget, measured in the number of tokens the model is allowed to use for internal computation. As shown in Figure 4, increasing this budget allows the model to scale its performance and achieve significantly higher accuracy.

2.6. Capability-specific improvements

While most of the changes made to our training architecture and recipe since Gemini 1.5 have resulted in improvements across all capabilities, we have also made changes that have resulted in some capability-specific wins. We will now discuss these for code, factuality, long context, multilinguality, audio, video, and agentic use cases (with a particular focus on Gemini Deep Research).

Code

Gemini 2.0 and 2.5 represent a strategic shift of our development priorities towards delivering tangible real-world value, empowering users to address practical challenges and achieve development objectives within today's complex, multimodal software environments. To realize this, concerted efforts have been undertaken across both pre-training and post-training phases since Gemini 1.5. In pre-training, we intensified our focus on incorporating a greater volume and diversity of code data from both repository and web sources into the training mixture. This has rapidly expanded coverage and enabled the development of more compute-efficient models. Furthermore, we have substantially enhanced our suite of evaluation metrics for assessing code capabilities aligned with downstream use cases, alongside improving our ability to accurately predict model performance.